Idiopathic Pulmonary Fibrosis

Radiology

Idiopathic Pulmonary Fibrosis:

Data-driven Textural Analysis of Extent of Fibrosis at Baseline and 15-Month Follow-up[1]

Stephen M. Humphries, PhD
Kunihiro Yagihashi, MD
Jason Huckleberry, MD
Byung-Hak Rho, MD
Joyce D. Schroeder, MD
Matthew Strand, PhD
Marvin I. Schwarz, MD
Kevin R. Flaherty, MD
Ella A. Kazerooni, MD
Edwin J. R. van Beek, MD
David A. Lynch, MB

Purpose: To evaluate associations between pulmonary function and both quantitative analysis and visual assessment of thin-section computed tomography (CT) images at baseline and at 15-month follow-up in subjects with idiopathic pulmonary fibrosis (IPF).

Materials and Methods: This retrospective analysis of preexisting anonymized data, collected prospectively between 2007 and 2013 in a HIPAA-compliant study, was exempt from additional institutional review board approval. The extent of lung fibrosis at baseline inspiratory chest CT in 280 subjects enrolled in the IPF Network was evaluated. Visual analysis was performed by using a semiquantitative scoring system. Computer-based quantitative analysis included CT histogram-based measurements and a data-driven textural analysis (DTA). Follow-up CT images in 72 of these subjects were also analyzed. Univariate comparisons were performed by using Spearman rank correlation. Multivariate and longitudinal analyses were performed by using a linear mixed model approach, in which models were compared by using asymptotic χ^2 tests.

Results: At baseline, all CT-derived measures showed moderate significant correlation ($P < .001$) with pulmonary function. At follow-up CT, changes in DTA scores showed significant correlation with changes in both forced vital capacity percentage predicted ($\rho = -0.41$, $P < .001$) and diffusing capacity for carbon monoxide percentage predicted ($\rho = -0.40$, $P < .001$). Asymptotic χ^2 tests showed that inclusion of DTA score significantly improved fit of both baseline and longitudinal linear mixed models in the prediction of pulmonary function ($P < .001$ for both).

Conclusion: When compared with semiquantitative visual assessment and CT histogram-based measurements, DTA score provides additional information that can be used to predict diminished function. Automatic quantification of lung fibrosis at CT yields an index of severity that correlates with visual assessment and functional change in subjects with IPF.

[1] From the Department of Radiology (S.M.H., D.A.L.) and Division of Biostatistics and Bioinformatics (M.S.), National Jewish Health, 1440 Jackson St, Denver, CO 80206-2761; Department of Radiology, St. Marianna University School of Medicine, Kawasaki, Japan (K.Y.); Department of Radiology, Kaiser Permanente, Denver, Colo (J.H.); Department of Radiology, Keimyung University School of Medicine, Daegu, Republic of Korea (B.H.R.); Department of Radiology and Imaging Sciences, University of Utah, Salt Lake City, Utah (J.D.S.); Division of Pulmonary Sciences & Critical Care Medicine, University of Colorado, Aurora, Colo (M.I.S.); Departments of Pulmonology (K.R.F.) and Radiology (E.A.K.), University of Michigan, Ann Arbor, Mich; and Clinical Research Imaging Centre, University of Edinburgh, Edinburgh, Scotland (E.J.R.v.B.). Received June 8, 2016; revision requested August 5; revision received January 2, 2017; accepted February 6; final version accepted February 23. **Address correspondence to** S.H. (e-mail: *humphriess@njhealth.org*).

Supported by the National Institutes of Health and the National Heart, Lung, and Blood Institute (R01 HL091743).

Idiopathic pulmonary fibrosis (IPF) is characterized by progressive scarring of the lungs and declining lung function, and it has a variable clinical course. Some patients experience periods of relative stability, while others have acute progression and rapid decline. Accurate diagnosis and prediction of disease course are difficult, prompting interest in standardized quantitative measures of severity to assist in clinical decision making and assessment of treatment efficacy in clinical trials (1,2). In clinical practice, disease severity is assessed with physiology, most commonly by using pulmonary function tests (PFTs), including measurement of forced vital capacity (FVC) and lung diffusing capacity for carbon monoxide (DLco), which are also the most common primary outcome variables used to assess progression in clinical trials. Known limitations of FVC and DLco in this context include lack of sensitivity to change and difficulty interpreting marginal changes of 5%–10% (2), emphasizing the need for more precise anatomically based measures of severity.

Extent of lung fibrosis on computed tomographic (CT) images is a recognized marker of disease severity (3); correlates with prognosis, including mortality (4); and can replace DLco in a clinical model to determine risk of death (1). The ability to evaluate the extent of fibrosis noninvasively, capturing the spatial and temporal heterogeneity typical of usual interstitial pneumonia (UIP), makes CT particularly useful. However, visual identification of a UIP pattern is impaired by interobserver variation, and results are generally communicated in nonspecific qualitative terms (5). Computer-based quantitative methods capitalize on detailed information available on images and have the potential to increase precision and reproducibility in evaluation of IPF on CT images.

The relatively recent approval of drug therapies for IPF has generated excitement that better treatments are within reach (6). As new treatments are evaluated, there is a need for an objective anatomy-based measure of IPF that could be provided by quantitative CT. Such a technique would be valuable in clinical practice to assess prognosis, identify progression, and standardize assessment across a diverse group of radiologists. In clinical trials, quantitative CT could be useful in the stratification of baseline severity of disease and in the assessment of progression or improvement in response to treatment.

Our group has developed a computer algorithm, dubbed data-driven textural analysis (DTA), which is capable of quantifying the extent of lung fibrosis on CT images. On the basis of initial technical validation (7), we hypothesized that serial DTA measurements are predictive of progression in patients with IPF. The purpose of this work was to evaluate associations between pulmonary function and both quantitative analysis and visual assessment of thin-section CT images at baseline and 15-month follow-up in patients with IPF.

Advances in Knowledge

- Extent of lung fibrosis determined with data-driven textural analysis (DTA) is associated with forced vital capacity (FVC) percentage predicted ($\rho = -0.60$, $P < .001$) and lung diffusing capacity for carbon monoxide (DLco) percentage predicted ($\rho = -0.68$, $P < .001$) at baseline.
- Change in DTA score is correlated with change in FVC percentage predicted ($\rho = -0.41$, $P < .001$) and DLco percentage predicted ($\rho = -0.40$, $P < .001$) at 15-month follow-up.
- Subjects with disease progression ($n = 34$) at visual assessment had greater mean DTA score at baseline (32.93 vs 23.43, $P = .003$), poorer function (mean FVC percentage predicted: 69.82 vs 78.34, $P = .019$; mean DLco percentage predicted: 41.91 vs 51.1, $P < .001$), and greater rates of change in these indexes over the follow-up period (change in DTA score: 6.6 vs 0.7, $P < .001$; change in FVC percentage predicted: −5.15 vs −1.27, $P = .003$; change in DLco percentage predicted: −5.59 vs −1.65, $P = .004$) when compared with patients who did not have disease progression.

Implication for Patient Care

- Automatic quantification of lung fibrosis on CT images yields an index of severity that correlates with visual assessment and functional change in subjects with idiopathic pulmonary fibrosis.

Materials and Methods

The original prospective study (8–10) was compliant with the Health Insurance Portability and Accountability Act and was approved by the institutional review board at all participating institutions; informed consent was obtained from all patients. As an analysis of preexisting anonymized data, this retrospective study of prospectively collected data was exempt from additional institutional review board approval.

https://doi.org/10.1148/radiol.2017161177

Content code: CH

Radiology 2017; 285:270–278

Abbreviations:
DLco = lung diffusing capacity for carbon monoxide
DTA = data-driven textural analysis
FVC = forced vital capacity
IPF = idiopathic pulmonary fibrosis
PFT = pulmonary function test
ROI = region of interest
SVM = support vector machine
UIP = usual interstitial pneumonia

Author contributions:
Guarantors of integrity of entire study, S.M.H., M.I.S., D.A.L.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, S.M.H., B.H.R., K.R.F., E.A.K., E.J.R.v.B., D.A.L.; clinical studies, K.Y., J.H., B.H.R., M.I.S., E.A.K., E.J.R.v.B., D.A.L.; statistical analysis, S.M.H., M.S.; and manuscript editing, S.M.H., K.Y., J.H., M.S., M.I.S., K.R.F., E.A.K., E.J.R.v.B., D.A.L.

Conflicts of interest are listed at the end of this article.

Subjects

For inclusion in the original studies, patients were required to have received a diagnosis of IPF according to standardized criteria similar to the subsequently published American Thoracic Society–European Respiratory Society criteria (11). Baseline CT examinations were performed within 3 months of enrollment. The minimal CT acquisition requirements for study entry were 0.5–1.25-mm thick sections obtained at 1-cm intervals and at full inspiration. When feasible, subjects underwent multidetector thin-section CT according to a prespecified protocol, which was customized according to CT scanner manufacturer and model. PFTs to measure FVC and DLco were performed according to standardized protocols (8–10). Data were acquired between 2007 and 2013.

CT studies that met minimal acquisition requirements were available in 539 subjects enrolled in the STEP, PANTHER, and ACE IPFnet trials (8–10). For this work, we selected subjects (n = 335) based on availability of thin-section volumetric CT (image reconstructions ≤ 1.25 mm) and PFT data. From this group, we selected a subset from the ACE cohort for algorithm training data (n = 55; mean age, 65.8 years; age range, 50–78 years [38 men; mean age, 65.3 years; age range, 50–78 years; 17 women; mean age, 67.1 years; age range, 54–75 years]). These subjects were selected for training because they had been examined with equipment from the most commonly used manufacturers (GE Healthcare, Waukesha, Wis; Siemens Healthineers, Erlangen, Germany) and because they had appropriate examples of the fibrotic UIP pattern. Twenty-two of these studies were acquired with Siemens equipment, and 33 were obtained with GE units. The remaining subjects with thin-section volumetric CT studies comprise the study subjects for this project (n = 280; mean age, 67.5 years; age range, 43–86 years [216 men; mean age, 67.9 years; age range, 48–86 years; 64 women; mean age, 66.3 years; age range, 43–84 years]). A total of 139 study subjects were examined with GE equipment, 89 were examined with Siemens equipment, and 52 were examined with Philips equipment.

Data in all subjects have been reported previously (8–10); however, prior articles dealt with experimental drug therapies, whereas we report a new method to quantify lung fibrosis on CT images.

Quantitative CT Analysis

DTA is based on an unsupervised feature-learning paradigm. Briefly, this method uses an initial clustering analysis to "learn" feature representations directly from a large collection of unlabeled images (7). Unsupervised clustering produces a set of basis elements, called a dictionary, which can be used to encode other image regions. Dictionary elements tend to be simple features, like directed edges or blobs, and they can be considered a collection of the elemental low-level pixel patterns that occur at lung CT. Weighting coefficients on dictionary elements needed to reconstruct novel image regions provide quantitative features useful for subsequent classification with machine learning. Unsupervised feature learning differs from feature engineering, which is a more traditional approach in which features are designed manually, often by using combinations of standard statistical or image processing calculations.

In our implementation, feature representations were produced by clustering thousands of small (3 × 3-mm) patches randomly sampled from lung regions on CT images in patients with IPF and in those without IPF. Images of the 55 subjects from the IPFnet ACE study, none of whom were included in this study cohort, served as examples of IPF. Images in nonsmoking control subjects from the Genetic Epidemiology of COPD (COPDGene) cohort (n = 35) (12) were used as examples of non-IPF lungs. A modified k-means clustering process (13) on the pixel intensities in these small patches yielded a 512-element dictionary of low-level CT patterns found in the lungs.

Separately, an experienced radiologist (K.Y., 10 years of experience) delineated regions of interest (ROIs) showing the characteristic patterns of UIP (honeycombing, reticular abnormality, and traction bronchiectasis) on CT images in 55 subjects from the IPFnet ACE cohort that were used in dictionary construction. A senior thoracic radiologist (D.L., more than 20 years of experience) verified each ROI. Quantitative features were computed for 14 × 14 mm ROIs by using the previously constructed dictionary. The features for a given 14 × 14 mm ROI can be thought of as the frequency with which 3 × 3 mm subregions matching learned dictionary elements occur in that ROI. These labeled exemplars were used to train a support vector machine (SVM) classifier. The SVM was trained in a binary fashion so that novel image regions described in terms of the established dictionary are classified as either a normal lung or a fibrotic lung (7). Hard negative mining, an iterative approach to training that emphasizes difficult-to-classify normal samples to reduce false-positive results, was used (14). Images of the 280 study subjects were not used to construct the dictionary or train the SVM classifier.

Analysis of study population CT studies began with semiautomatic lung segmentation with commercially available software (Apollo; Vida Diagnostics, Coralville, Iowa). Pixel histogram statistics (mean, standard deviation, skewness, and kurtosis) were computed over whole-lung volumes by using custom software developed in-house. Square 14 × 14 mm ROIs were sampled from within lung segmentation volumes and were classified with DTA as fibrotic or not fibrotic. DTA fibrosis score was computed as the number of ROIs classified as fibrotic divided by the total number of ROIs sampled from the lung segmentation volume.

Baseline CT Analysis

Volumetric CT of 280 subjects enrolled in ACE, PANTHER, and STEP IPFnet studies were used for testing (8–10). Two radiologists (J.H., K.Y.; 5 and 10 years of experience, respectively) scored extent of fibrosis, expressing the visual extent of fibrotic abnormality to the nearest 10% on a scale from 0 (no fibrosis) to 10 (100% fibrosis) (15). Baseline PFT data (FVC, FVC percentage predicted, DLco, and DLco percentage predicted) were available in all subjects.

Serial CT Evaluation

A subset of 72 subjects (51 men; mean age, 66.9 years; age range, 49–82 years; 21 women; mean age, 65.8 years; age range, 43–84 years) enrolled in the PANTHER study underwent both baseline and follow-up volumetric CT. The resultant images were evaluated independently by three senior thoracic radiologists (D.L., E.K., E.v.B.; all with more than 15 years of experience). Semiquantitative scores (scale, 0–10) for extent of fibrosis were assigned for each study. Visual assessment of change in fibrosis at follow-up was performed by comparing both studies and was scored on a five-point ordinal scale (0, much better; 1, slightly better; 2, same; 3, slightly worse; 4, much worse). Baseline and follow-up PFT results were available for all 72 subjects. FVC and FVC percentage predicted were measured at baseline and at 15, 30, 45, and 60 weeks thereafter. DLco and DLco percentage predicted were measured at baseline and at 30 and 60 weeks thereafter.

Table 1

Study Subject Characteristics

Characteristic	Baseline IPFNet (ACE, PANTHER, and STEP) ($n = 280$)	PANTHER ($n = 72$)		
		Baseline	Change at Follow-up	*P* Value*
Age (y)	67.5 ± 8.6	66.6 ± 9.2	…	<.001
Visual extent of fibrosis score†	4.24 ± 1.5	1.7 ± 0.7	0.3 ± 0.5	<.001
Visual assessment of change score‡	…	…	2.5 ± 0.6	<.001
DTA fibrosis score	36.6 ± 16.1	27.9 ± 13.4	4.0 ± 6.7	<.001
Mean lung attenuation (HU)	−714.0 ± 65.5	−728.9 ± 53.7	9.5 ± 32.6	.020
Standard deviation of lung attenuation (HU)	223.1 ± 34.2	208.8 ± 27.2	1.2 ± 15.8	.520
Skewness	1.7 ± 0.5	1.9 ± 0.5	−0.1 ± 0.2	<.001
Kurtosis	3.1 ± 2.6	3.9 ± 2.3	−0.4 ± 1.1	.002
CT total lung capacity (L)	3.7 ± 1.1	3.9 ± 1.0	−0.1 ± 0.4	.230
FVC (L)	2.7 ± 0.9	2.9 ± 0.7	−0.1 ± 0.3	<.001
FVC percentage predicted	66.2 ± 17.5	74.3 ± 15.4	−3.4 ± 6.5	<.001
DLco (mL of CO/min/mm Hg)	11.3 ± 4.4	13.8 ± 4.2	−1.2 ± 2.2	<.001
DLco percentage predicted	38.0 ± 13.4	46.8 ± 132.1	−4.3 ± 7.1	<.001

* *P* values were calculated with the one-sample *t* test.

† Semiquantitative extent of fibrosis was scored on a scale from 0 to 10. For subjects in the IPFnet study, scores at baseline were assigned by two reviewers (K.Y., J.H.). For patients in the PANTHER study, scores at baseline and follow-up were assigned by three reviewers (D.A.L., E.J.R.v.B., E.A.K.), and change at follow-up was calculated as follow-up score minus baseline score.

‡ Average of three readers (D.A.L., E.J.R.v.B., E.A.K.); a follow-up score of 0 indicated fibrosis was much better; a score of 1, fibrosis was slightly better; a score of 2, fibrosis was the same; a score of 3, fibrosis was worse; and a score of 4, fibrosis was much worse.

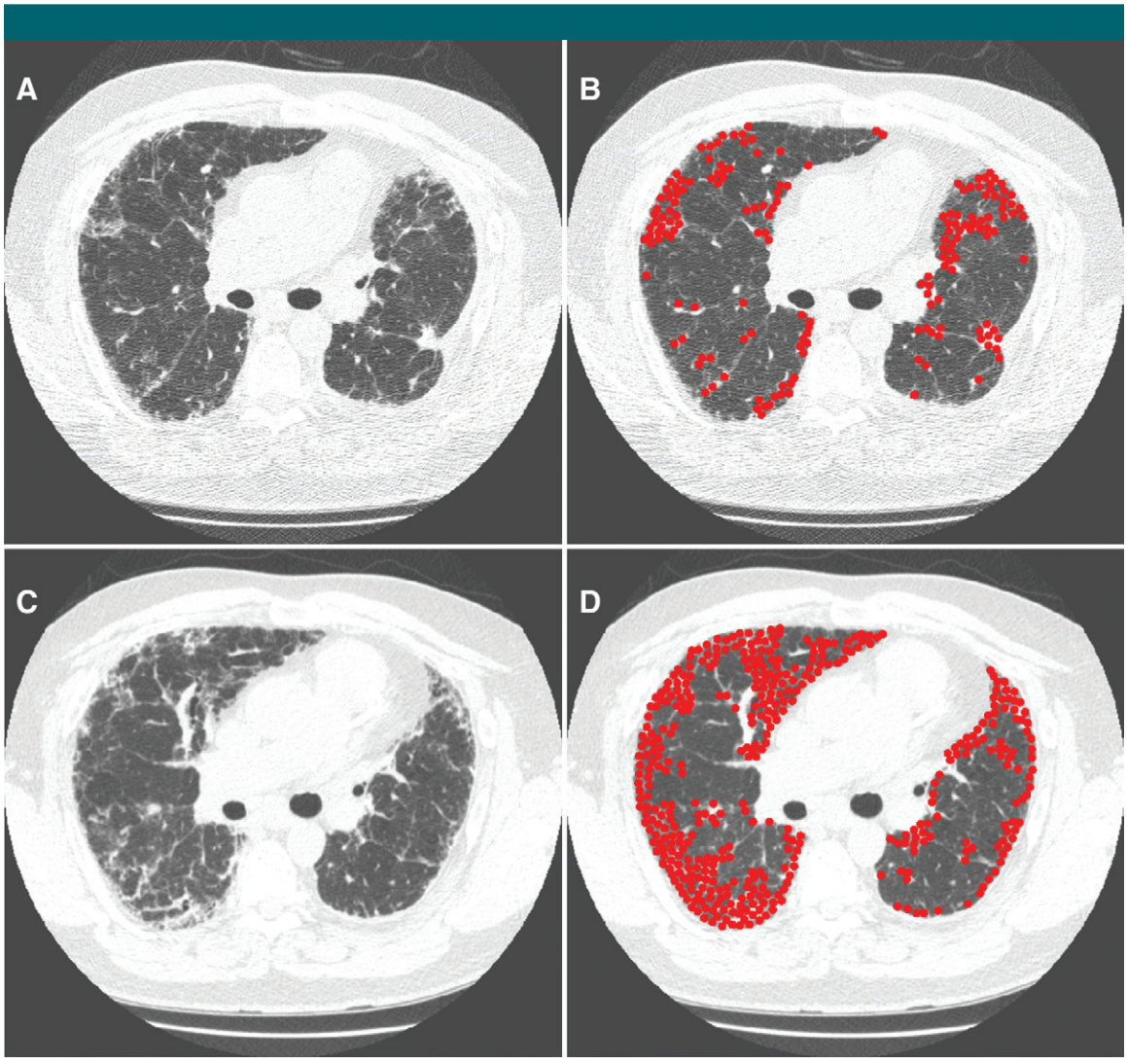
Statistical Analyses

Statistical analyses were performed by using two software packages (R, version 3.0.0, R Foundation for Statistical Computing, Vienna, Austria; SAS, version 9.4, SAS Institute, Cary, NC). Level of significance was set to $P < .05$ for all experiments, and correction for multiple comparisons was not applied. Spearman correlation coefficients were used to quantify univariate relationships between DTA scores, CT histogram-based measures (mean lung attenuation, standard deviation of lung pixel intensities, skewness, and kurtosis), semiquantitative visual scores, and physiologic measures (FVC, FVC percentage predicted, DLco, and DLco percentage predicted). Follow-up data were analyzed descriptively with Spearman correlation coefficients by considering change scores (last follow-up measurement minus baseline measurement) and rate of change (slope determined by linear regression of all follow-up measurements). For visual assessments, the mean of reviewers' scores was used in all analyses. Interrater agreement was evaluated with weighted κ scores. The κ agreement scores were interpreted by using the following criteria: κ of 0.0–0.2 indicated slight agreement; κ of 0.21–0.40, fair agreement; κ of 0.41–0.60, moderate agreement; κ of 0.61–0.80, substantial agreement; and κ of 0.81–1.0, almost perfect agreement (16). Subjects whose follow-up CT images were scored as slightly worse or much worse by at least two reviewers were considered to have progressed. Welch two-sample *t* tests were performed to compare baseline scores and rates of change between subjects whose disease had progressed and those whose disease had not.

Linear mixed models were used to evaluate multiple CT-derived approaches in predicting FVC or DLco for both baseline and longitudinal data. In particular, histogram, DTA, and visual-based variables were included as predictors in FVC or DLco models separately or together to determine their predictive utility or added benefit. Asymptotic χ^2 tests using the difference in two log likelihoods between nested models were used to determine whether DTA fibrosis score significantly improved prediction of clinical outcomes (17). Base models used visual score, CT histogram-based measures (mean, standard deviation, skewness, and kurtosis of lung voxel values), or both. Models for baseline data included one term per predictor (1 for visual score, 4 for CT histogram-based measures, 1 for DTA fibrosis score), while longitudinal models included two terms for each predictor (variable itself plus time-by-variable interaction). Thus, asymptotic tests for added benefit of DTA score used one degree of freedom in models for baseline data and two degrees of freedom in longitudinal models. Linear mixed models included a random intercept term for CT reconstruction kernel plus random terms for subject (in longitudinal models) and reader (in models for visual extent).

Results

Table 1 summarizes subject demographics. A subset of 72 subjects from



A, Axial baseline unenhanced CT image in a 60-year-old man. *B,* Areas of fibrosis identified with DTA are indicated (red ●). Baseline DTA fibrosis score was 22.3%, average visual extent of fibrosis as assessed by three readers (D.L., E.v.B., E.K.) was 1.67, baseline FVC percentage predicted was 53.15, and DLco percentage predicted was 36.1. *C,* Similar axial unenhanced CT image in the same subject at 18.2-month follow-up. *D,* DTA fibrosis score was 40.6% at follow-up, average visual extent of fibrosis (D.L., E.v.B., and E.K.) was 3.67, with all three readers scoring the follow-up image as slightly worse. Follow-up FVC percentage predicted was 44.4 and DLco percentage predicted was 30.5.

the PANTHER study underwent volumetric CT and PFT both at baseline and at follow-up. The follow-up interval ranged from 13.1 to 18.2 months (mean, 14.9 months ± 0.93 [standard deviation]; median, 14.7 months). The Figure shows visualization of DTA analysis results for baseline and follow-up images.

Table 2 shows associations between baseline values and rates of change of CT-derived measures, visual assessment, and PFTs. Radiologist agreement for semiquantitative fibrosis scores at baseline (assigned by two radiologists) was poor by weighted κ (κ = 0.3, $P < .001$, $n = 280$). However, DTA score showed moderate correlation with average visual score (Spearman ρ = 0.50, $P < .001$).

In the 72 subjects who underwent follow-up CT, the evaluation of agreement between three reviewers' visual assessment of change showed poor agreement (mean of pairwise weighted κ = 0.3, $P < .05$, $n = 72$). Agreement for extent of fibrosis scores at baseline and at follow-up (assigned by the same three readers) was moderate (mean pairwise κ = 0.53, $P < .001$). Subjects identified at visual assessment as having disease progression had higher DTA score and poorer PFT at baseline and had greater rates of change in these measurements over the follow-up period (Table 3).

Table 2

Associations between Baseline Values and Rates of Change for Visual Scores, Quantitative CT Measures, and Pulmonary Function

Characteristic	Baseline IPFNet (ACE, PANTHER, and STEP) ($n = 280$)					PANTHER ($n = 72$)				
	Visual Extent of Fibrosis Score*	FVC	FVC Percentage Predicted	DLco	DLco Percentage Predicted	Visual Follow-up Score	Change in FVC	Change in FVC Percentage Predicted	Change in DLco	Change in DLco Percentage Predicted
Visual extent of fibrosis score*	…	−0.22	−0.26	−0.46	−0.49	…	…	…	…	…
Visual assessment of change score†	…	…	…	…	…	…	−0.26 ($P = .030$)	−0.41	−0.41	−0.38
DTA fibrosis score	0.5	−0.54	−0.60	−0.64	−0.68	0.63	−0.42	−0.41	−0.42	−0.40
Mean (HU)	0.31	−0.62	−0.68	−0.54	−0.56	0.55	−0.38	−0.37	−0.32 ($P = .007$)	−0.28 ($P = .020$)
Standard deviation (HU)	0.45	−0.48	−0.55	−0.42	−0.45	0.30 ($P = .010$)	−0.52	−0.49	−0.07 ($P = .560$)	−0.03 ($P = .810$)
Skewness	−0.48	0.62	0.70	0.64	0.67	−0.49	0.47	0.44	0.24 ($P = .040$)	0.2 ($P = .080$)
Kurtosis	−0.49	0.62	0.68	0.62	0.65	−0.39	0.46	0.45	0.18 ($P = .140$)	0.16 ($P = .190$)
CT TLC (L)	−0.06 ($P = .360$)	0.83	0.57	0.49	0.31	−0.32	0.38	0.40	0.08 ($P = .500$)	0.08 ($P = .520$)

Note.—$P < .001$, except where indicated. Change at follow-up was computed as follow-up value minus baseline value.

* Average of two readers (K.Y., J.H.), where extent of fibrosis was scored as 0–10.

† Average of three readers (D.A.L., E.J.R.v.B., E.A.K.), with a visual follow-up score of 0 indicating fibrosis was much better; a score of 1, fibrosis was slightly better; a score of 2, fibrosis was the same; a score of 3, fibrosis was worse; and a score of 4, fibrosis was much worse.

CT-derived measurements were predictive of PFT outcomes for both baseline and longitudinal data based on linear mixed models. The main outcome variables analyzed (FVC, DLco) are approximately normally distributed based on examination of histograms. Table 4 shows the added benefit, based on asymptotic χ^2 tests, of including DTA score in models predicting FVC or DLco by using visual score, histogram measurement metrics, or both. The most significant improvements occurred when we included DTA score in models with visual score only ($P < .001$) and for all longitudinal models of DLco, including those with visual score, histogram measurements, or both ($P < .001$).

Discussion

We have developed a computer algorithm capable of quantifying extent of lung fibrosis on CT images. In 280 subjects with IPF, DTA fibrosis score showed correlation with semiquantitative visual scores and PFTs at baseline. In a subset of 72 subjects with baseline and follow-up CT images, DTA score showed moderate significant correlation with the degree of change in PFTs. This indicates that extent of lung fibrosis captured on CT images and quantified with DTA is associated with change in lung function. The purpose of comparing quantitative CT images with lung function is not to suggest that CT can or should replace PFTs. Our intent is to perform a clinical validation of CT analysis methods to show that measurement of morphology is associated with measures of disease severity; however, textural analysis may yield a more sensitive index of longitudinal change that may be in part independent of physiologic change. We focused on FVC and DLco because these measures are recommended for evaluation of disease course by the American Thoracic Society, European Respiratory Society, Japanese Respiratory Society, and Latin American Thoracic Association official guidelines on IPF (11). Multivariate analysis for prediction of lung function showed

DTA score adds information beyond histogram-based measurements and semiquantitative visual scores.

It is known that lung pixel histogram statistics, including skewness and kurtosis, correlate with baseline physiologic parameters, including FVC and DLco (4,17). Although these measures are attractive in that they are easy to calculate and are fairly intuitive, they are coarse global metrics that are not sufficiently precise for longitudinal analysis (4,18).

Automatic localization of interstitial lung disease on CT images is a challenging task. Fibrotic patterns are complex and are not defined adequately by using simple quantitative descriptors. Kliment et al (19) studied a large population and showed that attenuation value alone, separated by applying a threshold, is not sufficient to identify regions of interstitial lung abnormality. Local pixel histograms can yield more information. Iwasawa et al (20) described a method in which histograms of pixel intensities and gradients within ROIs were compared by using normalized correlation. They determined that the extent of a histogram pattern associated with honeycombing based on comparison with radiologist-identified exemplars was associated with mortality in a group of 40 patients with IPF. Bartholmai et al (21) and Maldonado et al (22) also used local histograms but classified small volumes of interest on thin-section CT images by using an approach based on clustering of expert-labeled volumes of interest by affinity propagation. By using this method to determine extents of UIP patterns on CT images, they reported correlations between automatic scores and other clinical measures in 119 subjects with interstitial lung disease.

Histograms count frequency of individual pixel intensities but do not explicitly capture local visually distinct patterns like edges or curves. The characteristic appearance of UIP on CT images is formed by patterns in both pixel intensity and spatial arrangement, a quality called image texture. Researchers have applied image texture methods to the problem of automatic analysis of interstitial lung disease on CT images and have shown impressive results (18,23,24). Kim et al (25) developed a method in which texture features computed by using local first- and second-order pixel statistics are classified with an SVM

Table 3

Comparison of Baseline Values and Change in Subjects Who Had Progressed and Those Who Had Not Based on 5% Decline in FVC Percentage Predicted and Visual Assessment

A: 5% Decline in FVC Percentage Predicted

Characteristic	Nonprogressors ($n = 42$)	Progressors ($n = 30$)	*P* Value*
Baseline			
DTA score	25.38 ± 11.21	28.41 ± 11.42	.273
DLco percentage predicted	47.72 ± 11.46	45.42 ± 12.97	.439
Visual extent†	1.70 ± 0.63	1.80 ± 0.74	.561
Rate of change (per year)			
DTA score	1.71 ± 5.0	6.92 ± 7.63	.002
DLco percentage predicted	−2.67 ± 6.96	−6.52 ± 6.84	.023
Visual assessment of change†	0.33 ± 0.61	0.822 ± 0.58	<.001

B: Visual Assessment

Characteristic	Nonprogressors ($n = 38$)	Progressors ($n = 34$)	*P* Value*
Baseline			
DTA score	23.43 ± 11.02	32.93 ± 14.16	.003
FVC percentage predicted	78.34 ± 13.53	69.82 ± 16.23	.020
DLco percentage predicted	51.10 ± 12.60	41.91 ± 9.49	<.001
Rate of change (per year)			
DTA score	0.70 ± 3.61	6.6 ± 5.68	<.001
FVC percentage predicted	−1.27 ± 4.94	−5.15 ± 5.56	.003
DLco percentage predicted	−1.65 ± 5.67	−5.59 ± 5.58	.004

Note.—Unless otherwise indicated, data are mean ± standard deviation.

* Welch two-sample *t* test.

† Average of three readers.

Table 4

***P* Values for Improvement in Goodness of Fit after Adding Algorithm Score as a Predictor in Clinical Outcome Models**

	Baseline Data		Longitudinal Data	
Adding DTA score to Model with the Following Base Predictors	FVC Percentage Predicted	DLco Percentage Predicted	FVC Percentage Predicted	DLco Percentage Predicted
Visual score	<.001 (81.8) [1]	<.001 (93.9) [1]	<.001 (17.2) [2]	<.001 (53.5) [2]
Histogram measures	.007 (7.2) [1]	.005 (7.9) [1]	.061 (5.6) [2]	<.001 (29.2) [2]
Histogram measures and visual score	.008 (7) [1]	.032 (4.6) [1]	.301 (2.4) [2]	<.001 (24.8) [2]

Note.—Data are *P* values. Data in parentheses are the χ^2 statistic, and data in brackets are degrees of freedom. Asymptotic χ^2 tests for differences in −2 log likelihood values were used to determine significance of improvement when adding algorithm score as a predictor in linear mixed models; see text for more detail.

trained in a supervised fashion using radiologist-delineated regions of lung fibrosis. In a study of 57 subjects with IPF and sequential CT examinations selected from a database of standardized scans, fibrosis scores computed with their method were associated with baseline FVC and DLco and change in these measurements after 7 months (18). Yoon et al (26) used a similar texture-based system to quantify disease patterns of fibrotic interstitial pneumonia at sequential CT in 89 subjects with clinically proven and/or biopsy proven UIP or nonspecific interstitial pneumonia. They showed there was an association between quantitative and visual scores, FVC and DLco at baseline, and change in FVC and DLco at 1 year follow-up. Correlations between DTA score and PFTs in the present study were similar to those reported by these investigators. Several recently developed classification systems have sought to subcategorize disease patterns into specific categories, including ground-glass abnormality, reticulation, and honeycombing (15,27). However, we chose to focus this analysis on the extent of fibrosis because in patients with IPF, the extent of fibrosis or reticular abnormality is invariably the most important in correlating with pulmonary function and outcome (28,29).

Most established image texture methods are based on engineered features, meaning calculation strategies designed to capture particular characteristics anticipated to be useful in distinguishing patterns of interest. These approaches can be limited in that they are based on assumptions regarding which characteristics are important, and they generally rely on multiple parameters that must be tuned, often by trial and error. The typical approach computes a battery of features then applies feature selection to identify those that are useful in the classification task. Unsupervised feature learning distills descriptive feature representations directly from a large collection of relevant data. Mounting evidence suggests this type of approach yields image features that are more powerful than engineered features (30), thereby providing more discriminative information for classifier training. This is the strategy used in DTA. A supervised classifier was trained by using radiologist-labeled regions, from which features were extracted by using a framework developed in an unsupervised process rather than by using designed features. There was no overlap in images used to develop the DTA algorithm, including construction of the feature dictionary and training the classifier, and the images analyzed in this study. Classification systems based on learned features are thought to be more flexible and easier to generalize to similar tasks (eg, classification of other CT patterns) than those based on engineered features (31).

Limitations of this study include variation in CT technical parameters, the relatively small number of subjects who underwent a follow-up examination, and the absence of mortality data. CT technical parameters, such as reconstruction kernel, affect image sharpness and noise level, which likely affect quantitative CT measures. The use of a random intercept term for reconstruction kernel in the linear mixed model accounts for this at multivariate analysis, but variation in CT technical parameters may contribute to the relatively low correlations at univariate analysis. The use of FVC and DLco as the reference standard for disease severity is inherently flawed because of known variation in these measures and because of the inevitable incongruity of physiologic and anatomic measurements. This may explain the relatively modest correlations between the CT-based and physiologic measurements. The lack of correction for multiple comparisons can also be considered a limitation; however, we chose to present individual P values for the planned comparisons. The observer variation for visual assessment is a known limitation that has been demonstrated in this study and in prior studies, but it emphasizes the difficulty of visual quantification and the need for an automated method. We also did not perform a direct comparison of DTA with other published fibrosis quantification methods. Nevertheless, a key conclusion from this and related work is that CT captures useful prognostic information that can be quantified by using computational techniques.

We conclude that DTA fibrosis score applied to CT from a multicenter study of IPF correlates with visual assessment and PFTs at baseline and when assessing change. Multivariate analysis indicates that DTA score improves the ability to predict change in function over time. Our findings that DTA fibrosis score is associated with baseline measurements of disease severity and with change over time suggest that CT captures anatomic information that can be quantified by using automatic computational methods to provide useful indexes of disease severity and progression in patients with IPF.

Disclosures of Conflicts of Interest: S.M.H. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: has a patent pending for a system and method for automatic detection and quantification of pathology using dynamic deature classification (62/335,816). Other relationships: disclosed no relevant relationships. **K.Y.** disclosed no relevant relationships. **J.H.** disclosed no relevant relationships. **B.H.R.** disclosed no relevant relationships. **J.D.S.** disclosed no relevant relationships. **M.S.** disclosed no relevant relationships. **M.I.S.** disclosed no relevant relationships. **K.R.F.** disclosed no relevant relationships. **E.A.K.** disclosed no relevant relationships. **E.J.R.v.B.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is the founder and owner of Quantitative Clinical Imaging Trails. Other relationships: disclosed no relevant relationships. **D.A.L.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: received personal fees from Paraxel, Genentech, and Boehringer Ingelheim; received a grant from Veracyte; has a patent pending for a system and method for automatic detection and quantification of pathology using dynamic deature classification (62/335,816). Other relationships: disclosed no relevant relationships.

References

1. Ley B, Elicker BM, Hartman TE, et al. Idiopathic pulmonary fibrosis: CT and risk of death. Radiology 2014;273(2):570–579.

2. Hansell DM, Goldin JG, King TE Jr, Lynch DA, Richeldi L, Wells AU. CT staging and monitoring of fibrotic interstitial lung diseases in clinical practice and treatment trials: a position paper from the Fleischner Society. Lancet Respir Med 2015;3(6):483–496.

3. Wells AU, Desai SR, Rubens MB, et al. Idiopathic pulmonary fibrosis: a composite physiologic index derived from disease extent observed by computed tomography. Am J Respir Crit Care Med 2003;167(7):962–969.

4. Best AC, Meng J, Lynch AM, et al. Idiopathic pulmonary fibrosis: physiologic tests, quantitative CT indexes, and CT visual scores as predictors of mortality. Radiology 2008;246(3):935–940.

5. Watadani T, Sakai F, Johkoh T, et al. Interobserver variability in the CT assessment of honeycombing in the lungs. Radiology 2013;266(3):936–944.

6. Hunninghake GM. A new hope for idiopathic pulmonary fibrosis. N Engl J Med 2014;370(22):2142–2143.

7. Humphries S, Yagihashi K, Sood R, Rho BH, Schroeder J, Lynch DA. Quantification of IPF on chest CT using unsupervised feature learning. ATS Journals Web site. http://www.atsjournals.org/doi/abs/10.1164/ajrccm-conference.2015.191.1_MeetingAbstracts.A4973. Accessed September 29, 2015.

8. Idiopathic Pulmonary Fibrosis Clinical Research Network, Zisman DA, Schwarz M, et al. A controlled trial of sildenafil in advanced idiopathic pulmonary fibrosis. N Engl J Med 2010;363(7):620–628.

9. Idiopathic Pulmonary Fibrosis Clinical Research Network, Raghu G, Anstrom KJ, King TE Jr, Lasky JA, Martinez FJ. Prednisone, azathioprine, and N-acetylcysteine for pulmonary fibrosis. N Engl J Med 2012;366(21):1968–1977.

10. Noth I, Anstrom KJ, Calvert SB, et al. A placebo-controlled randomized trial of warfarin in idiopathic pulmonary fibrosis. Am J Respir Crit Care Med 2012;186(1):88–95.

11. Raghu G, Collard HR, Egan JJ, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis—evidence-based guidelines for diagnosis and management. Am J Respir Crit Care Med 2011;183(6):788–824.

12. Zach JA, Newell JD Jr, Schroeder J, et al. Quantitative computed tomography of the lungs and airways in healthy nonsmoking adults. Invest Radiol 2012;47(10):596–602.

13. Coates A, Ng AY. Learning feature representations with k-means. In: Montavon G, Orr GB, Müller KR, eds. Neural networks: tricks of the trade. Berlin, Germany: Springer, 2012; 561–580.

14. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005. Piscataway, NJ: IEEE, 2005; 886–893.

15. Salisbury ML, Lynch DA, van Beek EJ, et al. Idiopathic pulmonary fibrosis: adaptive multiple features method fibrosis association with outcomes. Am J Respir Crit Care Med 2016 Oct 21. [Epub ahead of print]

16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–174.

17. Best AC, Lynch AM, Bozic CM, Miller D, Grunwald GK, Lynch DA. Quantitative CT indexes in idiopathic pulmonary fibrosis: relationship with physiologic impairment. Radiology 2003;228(2):407–414.

18. Kim HJ, Brown MS, Chong D, et al. Comparison of the quantitative CT imaging biomarkers of idiopathic pulmonary fibrosis at baseline and early change with an interval of 7 months. Acad Radiol 2015;22(1):70–80.

19. Kliment CR, Araki T, Doyle TJ, et al. A comparison of visual and quantitative methods to identify interstitial lung abnormalities. BMC Pulm Med 2015;15(1):134.

20. Iwasawa T, Asakura A, Sakai F, et al. Assessment of prognosis of patients with idiopathic pulmonary fibrosis by computer-aided analysis of CT images. J Thorac Imaging 2009;24(3):216–222.

21. Bartholmai BJ, Raghunath S, Karwoski RA, et al. Quantitative computed tomography imaging of interstitial lung diseases. J Thorac Imaging 2013;28(5):298–307.

22. Maldonado F, Moua T, Rajagopalan S, et al. Automated quantification of radiological patterns predicts survival in idiopathic pulmonary fibrosis. Eur Respir J 2014;43(1):204–212.

23. Uppaluri R, Hoffman EA, Sonka M, Hunninghake GW, McLennan G. Interstitial lung disease: a quantitative study using the adaptive multiple feature method. Am J Respir Crit Care Med 1999;159(2):519–525.

24. Kim HJ, Li G, Gjertson D, et al. Classification of parenchymal abnormality in scleroderma lung using a novel approach to denoise images collected via a multicenter study. Acad Radiol 2008;15(8):1004–1016.

25. Kim HG, Tashkin DP, Clements PJ, et al. A computer-aided diagnosis system for quantitative scoring of extent of lung fibrosis in scleroderma patients. Clin Exp Rheumatol 2010;28(5,Suppl 62):S26–S35.

26. Yoon RG, Seo JB, Kim N, et al. Quantitative assessment of change in regional disease patterns on serial HRCT of fibrotic interstitial pneumonia with texture-based automated quantification system. Eur Radiol 2013;23(3):692–701.

27. Jacob J, Bartholmai BJ, Rajagopalan S, et al. Automated quantitative computed tomography versus visual computed tomography scoring in idiopathic pulmonary fibrosis: validation against pulmonary function. J Thorac Imaging 2016;31(5):304–311.

28. Lynch DA, Godwin JD, Safrin S, et al. High-resolution computed tomography in idiopathic pulmonary fibrosis: diagnosis and prognosis. Am J Respir Crit Care Med 2005;172(4):488–493.

29. Park HJ, Lee SM, Song JW, et al. Texture-based automated quantitative assessment of regional patterns on initial CT in patients with idiopathic pulmonary fibrosis: relationship to decline in forced vital capacity. AJR Am J Roentgenol 2016;207(5):976–983.

30. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. IEEE Intell Syst 2009;24(2):8–12.

31. Coates A, Ng AY, Lee H. An analysis of single-layer networks in unsupervised feature learning. In: International Conference on Artificial Intelligence and Statistics, 2011; 215–23.